Reducing the length of mental health instruments trough structurally incomplete designs

Smits, N.; Cuijpers, P.; Beekman, A.T.F.; Smit, J.H.

published in
International Journal of Methods in Psychiatric Research
2007

DOI (link to publisher)
10.1002/mpr.223

document version
Publisher's PDF, also known as Version of record

International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 16(3): 150–160 (2007)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.223

Reducing the length of mental health instruments through structurally incomplete designs

NIELS SMITS,[1] PIM CUIJPERS,[1] AARTJAN T. F. BEEKMAN,[2] JOHANNES H. SMIT[2]

1 Department of Clinical Psychology, Vrije Universiteit, Amsterdam, The Netherlands
2 Institute for Research in Extramural Medicine, Vrije Universiteit Medical Centre, Amsterdam, The Netherlands

Abstract

This paper presents structurally incomplete designs as an approach to reduce the length of mental health tests. In structurally incomplete test designs, respondents only fill out a subset of the total item set. The scores on the unadministered items are estimated using methods for missing data. As an illustration, structurally incomplete test designs recording, respectively, two thirds, one half, one third and one quarter of the complete item set were applied to item scores on the Centre of Epidemiological Studies-Depression (CES-D) scale of the respondents in the Longitudinal Aging Study Amsterdam (LASA). The resulting unobserved item scores were estimated with the missing data method Data Augmentation. The complete and reconstructed data yielded very similar total scores and depression classifications. In contrast, the diagnostic accuracy of the incomplete designs decreased as the designs had more unobserved item scores. The discussion addresses the strengths and limitations of the application of incomplete designs in mental health research. Copyright © 2007 John Wiley & Sons, Ltd.

Key words: missing data, structurally incomplete designs, diagnostic accuracy, data augmentation, multiple imputation

Introduction

In mental health research, self report questionnaires are frequently used to assess mental health. Developers of measurement instruments often have to consider the length of their questionnaires. Long instruments may provide much information, but also take more testing time, increasing the responder's burden. This may have several disadvantageous outcomes. First, for some respondent groups, such as the elderly, the attention span is short and administering many questions may decrease the quality of the answers. Second, the willingness to participate and/or complete questionnaires may decrease (e.g. Dillman et al., 1993; Wacholder et al., 1994). In addition, faced with a fixed booklet size, longer tests do not allow for the collection of data on other variables.

One solution that has been proposed (e.g. Burisch, 1997) is to use shortened versions of instruments: a subset of best items is selected and administered. The suitability of items for selection is often assessed using classical psychometric criteria such as item-total correlations or item-criterion correlations (e.g. Jongenelis et al., 2005). A more principled method is Item Response Theory (IRT), in which it is assumed that a single construct such as depression underlies the items that make up a scale (e.g. Tang et al., 2005). For each item it is assessed how suitable it is for measuring the construct in question. Moreover, IRT allows for

computerized adaptive testing (CAT). CAT makes it possible to individualize each assessment so that only the most informative questions are asked of each person.

Although such approaches may be fruitful in many applications, they have at least two disadvantages. If a researcher wishes to shorten a scale herself, then she must have information of the psychometric properties of the items before research is conducted. Such information may not always be available. In addition, the known item properties may not be representative for the population to which the researcher wishes to administer the test. Second, only a part of the originally constructed items is administered. Consequently, the test may no longer completely provide data on a set of original variables. This may be undesirable when, for example, an instrument is designed to completely cover the symptomatology associated with a mental disorder.

An approach that does not suffer from these two drawbacks is to use structurally incomplete test designs. Using such designs, questionnaires are split into several item subsets and only some of the subsets are administered to any respondent (see, e.g. Wacholder et al., 1994; Raghunathan and Grizzle, 1995; Graham et al., 1996). Scores to the unadministered subsets are estimated using methods for missing data. As a result, data on all the items for all subjects are available. Of course, the application of these designs and the missing data method should yield sound estimates of the unavailable scores.

Structurally incomplete test designs have proved to be useful in educational and psychological testing (Kaplan, 1995; Mislevy and Wu, 1996). The reliability and validity of such tests did not suffer from the application of these designs. However, the usefulness of structurally incomplete designs has not been studied for mental health instruments. As such instruments are often used as first stage screener, the usefulness of applying incomplete designs to mental health scales should be assessed by studying their effect on diagnostic accuracy.

In general, data that are intentionally not recorded are called 'missing by design' or 'planned missing'. If the booklets of a test with a structurally incomplete design are distributed randomly (each respondent has the same probability of receiving a given booklet), then the resulting planned missing data have the most advantageous missing data mechanism. Biases should not exist because subjects who receive a given booklet are not expected to be systematically different from subjects who receive other booklets (Rubin, 1987). Therefore, these data are called missing completely at random. This type of missing data mechanism allows for the application of several advanced methods to deal with missing values.

In the last two decades a large number of methods for dealing with missing values has been developed (see, e.g. Smits et al., 2002). A distinction between naive and more principled methods can be made (Little and Schenker, 1995). In the literature (e.g. Bernaards and Sijtsma, 1999; Sijtsma and van der Ark, 2003; Schafer and Olsen, 1998) both types of approaches have been compared and evaluated. The two most advanced methods of dealing with missing values are Maximum Likelihood (ML) and Multiple Imputation (MI) (Enders, 2001; Schafer and Graham, 2002). The two methods have in common that missing data are predicted from observed data. MI actually replaces, or imputes, unobserved values in the data file with estimates. In contrast, ML estimates parameters from the data (such as means and a correlation matrix). Therefore, MI has the advantage that after the dealing with missing values data files without missing values can be analysed. Another advantage is that MI explicitly takes into account the extra uncertainty that is a natural result of missing values. A more extensive description of MI will be given in the Method section.

In practice, when applying structurally incomplete test designs, a part of the potentially complete data is not recorded. For these designs to be a good method to reduce the length of mental health tests, the estimated scores to the missing items should be similar to such unobserved values. Moreover, the analysis of the complete data and the analysis of the imputed data should yield similar inferences. In this article, it is studied whether the application of both structurally incomplete designs and MI does indeed yield inferences that are comparable to inferences based on complete test data. Particularly, the diagnostic accuracy of imputed mental health tests is examined.

Method

In this study, a large data file consisting of the scores on a mental health scale was used. Structurally incomplete test designs with different rates of missing values were applied to these data. On the basis of these designs, item responses were made unobservable. Next, a MI method was applied to impute the missing values.

DOI: 10.1002/mpr

Finally, it was studied how well the complete and imputed data were in line with each other.

Participants

The Longitudinal Aging Study Amsterdam (LASA) is a longitudinal study over 10 years on predictors and consequences of changes in well-being and autonomy in the population aged 55 to 85 years (Deeg et al., 1993). Sampling and procedures have been reported in detail elsewhere (see Beekman et al., 1994, 1995). At baseline (1992/1993), a large (N = 3107) random sample of inhabitants, drawn from 11 community registries in three regions of the Netherlands, was interviewed. All interviews were conducted in the homes of respondents by trained and intensively supervised interviewers. Informed consent was obtained prior to the study, in accordance with legal requirements in the Netherlands. From this sample, respondents were selected who had complete data on all of the Centre of Epidemiological Studies-Depression (CES-D) scale items. This resulted in a sample of N = 2852. Of these, 317 subjects were screen positives on the basis of CES-D. They, and a similarly sized random sample of the screen negatives (N = 307) received diagnostic interviews with the Diagnostic Interview Schedule (DIS) (see Diagnostic criterion section) to assess the presence of major depressive disorder.

Depressive Symptomatology

The CES-D (Radloff, 1977) is commonly used as a first-stage screener for depression, and as an indicator of the severity of depression. The CES-D is a 20-item scale. The items have a four point (0–3) Likert rating scale. The total score ranges from 0 to 60. It has been widely used in older community samples and has good psychometric properties in this age group (see, e.g. Himmelfarb and Murrell, 1983; Hertzog et al., 1990). The Dutch translation had similar psychometric properties in three previously studied samples of elderly in the Netherlands (Beekman et al., 1994). In order to identify respondents with clinically relevant levels of depression, the cut-off score ≥ 16 is commonly used. Using this score, good criterion validity of the CES-D for assessing depression has been reported (e.g. Beekman et al., 1997).

Diagnostic criterion

The DIS (Robins et al., 1991) is often used as a second-stage instrument after a positive indication on the CES-D. The DIS was designed for epidemiological research and has been widely used, also among the elderly. In the LASA project, interviewers were fully trained by certified staff, using the official Dutch translation of the DIS (Dingemans et al., 1985). Obviously, for the utility of the CES-D for screening depression, a good relationship with the DIS is required.

In the total sample 14.8% of the respondents scored above the cut-off on the CES-D. In the group selected for receiving the DIS, the ratio of positive to negative CES-D diagnoses was stratified to be approximately one (the actual result was 50.8% positive diagnoses, and 49.2% negative diagnoses). On the basis of this sample, the criterion validity of the CES-D was assessed by employing the DIS diagnosis as gold-standard. As stratification influences sensitivity and specificity, these percentages were weighted using appropriate weights (the ratio of subjects scoring < 16 to those scoring ≥ 16 is 85.2 : 14.8 = 5.76 : 1) to obtain estimates of diagnostic accuracy for the complete (N = 2852) sample. The resulting sensitivity, specificity, positive predictive value, and negative predictive value were, 70%, 87%, 17% and 99%, respectively.[1]

Applying the incomplete test design

When applying structurally incomplete test designs, a part of the potentially complete data is not recorded. Here, it is studied to what extent the estimates of missing data methods approach such unobserved values.

In this study the balanced incomplete block design (see Johnson, 1992) was used. The most simple version of this design results in three booklets. The item set is divided in three equal parts, A, B, and C. The first booklet contains item sets A and B, the second booklet contains item sets B and C, and the third booklet contains item sets A and C. As a result, for each respondent two-thirds of the potentially complete item set is administered, and one third is missing. The test design can be extended such that a smaller part of the full item set is administered. Such designs need more booklets, however. For a design in which only half the

[1]The diagnostic accuracy of the CES-D as assessed using the DIS is often reported to be higher (e.g. Beekman et al., 1997). However, the recency of the DIS diagnosis used in those studies (1 month) is lower than the recency in our study (6 months). Therefore, it is to be expected that the diagnostic accuracy in the present analysis is lower than in the other studies.

DOI: 10.1002/mpr

items is administered for each subject, six booklets are needed. The item set is then divided in four equal sets A, B, C, and D. The first booklet contains item sets A and B, the second booklet contains A and C, the third contains sets A and D, the fourth contains sets B and C, the fifth contains sets B and D, and the sixth booklet contains sets C and D. Likewise, designs that record only a third and a quarter of all item scores can be applied, which need, respectively, 15 and 28 booklets.

Compared to other designs, the balanced incomplete block design has some advantageous features. First, when test booklets are handed out randomly, the items are administered to approximately equal numbers of examinees. Second, all item pairs are administered in at least one booklet, allowing for a good estimation of inter-item correlations. Third, a relatively small number of booklets is needed. For a more extensive description of the utility of this design for the estimation of missing values, see for example, Graham et al. (1996), or Johnson (1992).

In this study, balanced incomplete block designs with four different rates of missing values were applied to the CES-D: a third (M1/3), a half (M1/2), two-thirds (M2/3) and three-quarters (M3/4). The incomplete test designs were applied to the data by deleting the item scores that were not to be observed according to the design. For the construction of the virtual booklets, the order of the items in the original instrument was preserved; the original 1st item was always in the first item set, and the 20th item was always in the last item set.

For two designs, M1/3 and M2/3, it was not possible to select subsets with equal numbers of items (because the number of items, 20, is not divisible by 3 and 6, respectively). In these designs, subsets with unequal numbers of items were used. For example, in the M1/3 design, the items of the CES-D were divided in such a way over the booklets that two booklets contained 13 items, and one booklet contained 14 items. The resulting incomplete data matrix had observed values for two-thirds of the participants, for all items.

Data augmentation

MI was performed using a procedure called Data Augmentation (DA), which is the most sophisticated method available to create MIs (Allison, 2001). It was carried out as described by Schafer (1997), and Schafer and Olsen (1998). DA predicts missing values using a regression-like procedure. It conforms roughly to the following procedure: from the available data the relationships between the variables are estimated. On the basis of these relationships, each variable can be regressed upon the other variables. For each observation with a missing value, the observed values are entered in a regression equation to produce a prediction.

The procedure is somewhat more intricate, however. MI was developed to deal with missing data in a way that is in accordance with common statistical inference. The problem of incomplete data is, of course, that data values intended to be observed are in fact missing. These missing values result in larger uncertainty (sampling error) in the outcomes because of the reduced size of the data base. To take this into account, imputations are randomly drawn from a predictive distribution of plausible values. So, rather than giving a specific prediction, the regression approach described earlier creates a distribution of predictions with a given mean and standard deviation out of which a value is randomly drawn. The mean of this distribution is the best prediction and the standard deviation is influenced by both the strength of the relationships among the variables and proportion of missing values. Of course, values closer to the mean have a larger probability of being drawn than values in the extremities of the predictive distribution. MI performs several draws from the predictive distributions resulting in multiple data sets that contain both the observed data and imputed values. Next, each of the data files is analysed in the normal way. The outcomes (e.g. a mean or a correlation coefficient) of each analysis are then combined to produce a single overall inference. As an estimate of a given outcome, the average outcome over the multiple analyses is calculated. The variability among the multiple outcomes provides a measure of the uncertainty with which the missing values are derived from the observed ones (Schafer and Olsen, 1998). Using rules provided by Rubin (1987), estimates of standard errors can be calculated and statistical tests can be performed (e.g. a mean deviating from zero). This paper is not intended as an introduction to MI, but as an illustration of the usefulness of applying MI on incomplete data to reduce test length. Here, we will primarily focus on the outcomes of the imputed data and the extent to which they approach the complete data outcomes. The calculation of proper standard errors and the carrying out of statistical tests will not be addressed here. For a more extensive description of MI and DA, see, for example, Schafer and Olsen (1998) or Schafer (1997).

DOI: 10.1002/mpr

DA procedures assuming different distributions of the missing values are available (see Schafer, 1997). In the present study, DA was performed using the NORM procedure (written by Schafer (1999), downloadable free of charge at http://www.stat.psu.edu/~jls/misoftwa.html). In NORM it is assumed that the missing data have a multivariate normal distribution. Commonly, item responses may at best be approximately normally distributed. However, NORM has proven to be quite robust against departures from the imputation model (Schafer and Olsen, 1998).

Each of the four data files (resulting from the four designs with different rates of missing values) was imputed five times. The imputed values were brought into line with the original rating scale: values were rounded to the nearest value on the scale (e.g. 3.45 to 3), and imputed values that fell out of the range of the rating scale were recoded to the closest minimum or maximum value (e.g. 4.30 to 4).

In this study, the CES-D total score of the subjects, and not some population quantity, was the parameter of main interest. For each subject, the best estimate was obtained by averaging over the total scores that were calculated in each of the five imputed data files. This estimate of the CES-D total score was subsequently used to compare complete and imputed data.

When applying DA, the predictive distribution from which imputations are drawn should be studied. DA is an iterative estimation procedure which should converge before data files may be imputed. To check this convergence, diagnostics were conducted in NORM as described by Schafer (1997, chapters 4 and 5).

Comparing complete and imputed data

Three outcomes were studied to determine the extent to which estimated item scores were in accordance with the original item scores: total scores, classifications, and CES-D × DIS classification tables.

Mental health tests are frequently used to position subjects on some scale: a higher score may represent a poorer mental health than a lower score. A score is commonly obtained by calculating the sum of the item scores. Therefore, the DAs ability to estimate total scores is studied. To that end, the correlation between complete data total scores and reconstructed data total scores are studied. In addition, because the two types of total scores share the observed item scores, the intra-class correlation coefficient was calculated as well. However, as the intra-class correlations were very similar to the Pearson correlations (maximal difference of two hundredths), results on these coefficients will not be reported.

In addition, mental health tests are used as screeners of mental disorders as well. A person receives a positive indication if the test score is above some cut-off point, and a negative indication when it is below this cut-off point. To study the effect of the incomplete test design on the CES-D diagnoses, the agreement between complete and incomplete data diagnoses was studied. To that end Cohen's (1960) kappa for the agreement of nominal classifications between two raters was used. In addition, to compare the proportion of subjects meeting the criteria for depression on the complete and imputed CES-D, McNemar's test for the difference between two correlated proportions (e.g., May and Johnson, 1997) was used.

The third outcome that was studied was the impact of incomplete test designs on the utility of the CES-D as a first-stage screener of depression. To that end, the relationship of imputed CES-D with second-stage instrument DIS was studied. On the basis of the agreement between complete and imputed CES-D, and the relationship between complete sample ($N = 2852$) CES-D and DIS (see, the Diagnostic criterion section), the relationship between imputed CES-D and DIS was derived. For each rate of missing values, the relationship between imputed CES-D and DIS is expressed in an Receiver Operating Characteristic (ROC) curve, classification table, sensitivity, specificity, positive predictive values and negative predictive values for the whole sample. More specifically, it was studied how many respondents suffering from depression would be classified as such after applying incomplete test designs to the CES-D.

Results

Multiple imputation of missing item scores

For all four rates of missing values, the distribution of predictive values of DA showed good convergence. For the two designs with the highest rate of missing values (M2/3 and M3/4), a minor modification in the default procedure in NORM was applied to stabilize the estimation of missing values (see Schafer, 1997, p. 156).[2]

[2]Detailed descriptions of the estimation and imputation processes in NORM are obtainable from the first author.

DOI: 10.1002/mpr

Total scores

The second column of Table 1 shows the correlations between the original data and imputed data total scores. The correlations were relatively high. For example, although the respondents answered only five of the 20 items in the M3/4 design, a correlation of 0.825 with the original 20 items total scores was found. Self evidently, the correlations between complete and imputed CES-D data became lower as the number of administered items decreased. In addition, the estimates of the reliability coefficients of the CES-D resulting from the imputed data were very close to the coefficients of the complete data CES-D; whenever different, they deviated only by a few hundredths.

Diagnostic agreement

The third column of Table 1 shows the 2 × 2 tables for the classification (depressive versus not depressive) of the complete and imputed CES-D. Complete and incomplete CES-D gave mostly identical classifications for all rates of missing values. Three of the four incomplete versions of the CES-D gave positive indications somewhat more often than the complete version of the CES-D. However, McNemar's test for the equality of correlated proportions showed that these differences were not statistically significant. Evidently, the degree of agreement diminished as the number of administered items decreased. The second column of Table 1 shows the kappa coefficients associated with the agreement in diagnosis of the imputed data with the complete data. Benchmarks for the classification of the strength of agreement associated with kappa are available (e.g. Landis and Koch, 1977). Although these classifications are somewhat arbitrary, they are practical guidelines when comparing the four incomplete designs. Kappa values ranging from 0.41 to 0.60 are said to represent moderate agreement, values from 0.61 to 0.80 represent substantial agreement, and values higher than 0.80 to 1.00 represent almost perfect agreement. For all rates of missing values, the agreement of imputed data diagnosis with the complete data diagnosis was at least moderate. When a third part of the data was missing, there was an almost perfect agreement with the complete data diagnosis (kappa = 0.844). This agreement became somewhat lower as the rate of missing values increased: substantial agreement for both a half and two-thirds missingness (kappas of 0.778, and 0.679, respectively), and moderate agreement for three-quarters missingness (kappa = 0.606).

Table 1. Association between complete and imputed CES-D as a function of proportion of missingness

Proportion missing	Correlations	Kappa	Classification table			
					ICES-D	
					0	1
1/3	0.964	0.844	CES-D	0	2 381	50
				1	61	360
					ICES-D	
					0	1
1/2	0.912	0.778	CES-D	0	2 363	68
				1	88	333
					ICES-D	
					0	1
2/3	0.873	0.679	CES-D	0	2 303	128
				1	107	314
					ICES-D	
					0	1
3/4	0.825	0.606	CES-D	0	2 292	139
				1	143	278

Note: ICES-D is the incomplete and imputed CES-D.

DOI: 10.1002/mpr

Criterion validity

Figure 1 shows ROC curves associated with the complete and incomplete CES-D. As can be seen in Figure 1, on average, designs that record scores on fewer items have worse ROC curves. In addition, Table 2 is associated with the validity of the incomplete design versions of the CES-D as a first stage screener of depression. Column two depicts the classification table for each rate of missing values. The diagnostic accuracy clearly drops: as the number of observed items decreases, the number of misclassifications (the numbers in the off-diagonal cells) increases. This bias is most apparent in the sensitivity (column three) and positive predictive value (column five). As the rate of missing values increased from 0 to 3/4, a drop of 25 percentage points, and a drop of 6 percentage points, were found, respectively.

In contrast, the specificity (column four) and negative predictive values (column six) hardly suffered from the application of incomplete designs. The negative predictive value shows a drop of only 1 percentage

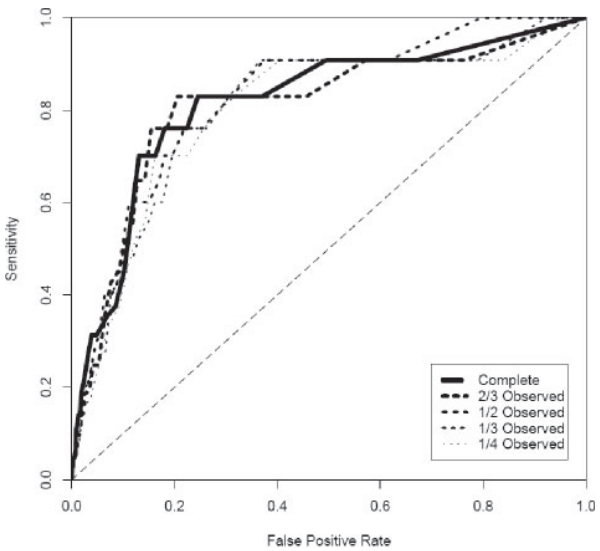

Figure 1. ROC curves of the complete (solid line) and incomplete/imputed CES-D (dashed lines). Thinner lines represent structurally incomplete designs with lower rates of observed item scores.

Table 2. Diagnostic accuracy as a function of proportion of missingness

Proportion missing	Classification table				Sensitivity	Specificity	Positive predictive value	Negative predictive value
			DIS					
			0	1				
0	CES-D	0	2 399	31	0.699	0.874	0.173	0.987
		1	349	72				
			DIS					
			0	1				
1/3	CES-D	0	2 401	41	0.602	0.873	0.151	0.983
		1	348	62				
			DIS					
			0	1				
1/2	CES-D	0	2 405	46	0.558	0.872	0.145	0.981
		1	343	58				
			DIS					
			0	1				
2/3	CES-D	0	2 362	48	0.534	0.859	0.124	0.980
		1	387	55				
			DIS					
			0	1				
3/4	CES-D	0	2 381	54	0.447	0.866	0.118	0.978
		1	368	49				

DOI: 10.1002/mpr

point when instead of 20, only five items were administered. The specificity only showed a drop of 2 percentage points.

For each of the four measures of diagnostic accuracy, the outcomes of the imputed data showed a high relationship with the rate of missing values. For example, the predictive positive value was almost perfectly related to the rate of missing values ($r = -0.99$, $p < 0.001$). This suggests that the diagnostic accuracy can be predicted from the rate of missing values.

The practical impact of planned missing values on diagnostic accuracy may best be illustrated when studying the number of correct positives. This number is shown in the lower right cell of each classification table. The number of depressive respondents screened positive decreased as the rate of missing values went from 1/3 to 3/4 (62, 58, 55, and 49, respectively). Of the complete data correct positives (72 respondents), respectively, 13.8% ($100 \times [72 - 62]/72$), 19.4%, 23.6%, and 31.9% were false negatives in the incomplete CES-D administrations (in M1/3, M1/2, M2/3, and M3/4, respectively). Although these rates are proportional to the rates of missing values, they are much lower in absolute value. In other words, the proportions of depressed respondents who were detected in the complete data, but who were 'missed' in the incomplete versions of CES-D were much lower than the matching rates of missing values.

Discussion

In this study it was shown how structurally incomplete test designs can be applied as a method of reducing the length of mental health tests. It should be noted that these designs are especially useful in research settings. In contrast, in clinical settings, the administration of complete instruments is needed.

MI method DA yielded sound estimates of the scores on the unadministered items that resulted from the incomplete designs. A high correspondence between complete and imputed data CES-D total scores, and between complete and imputed data depression classifications was found for all rates of unobserved item scores. In addition, the complete and imputed CES-D did not significantly differ in the proportion of positive classifications. Self-evidently, in spite of a high agreement between complete and imputed CES-D scores in all four designs, the agreement between complete and imputed data decreased as the designs had a lower number of observed values. In contrast, when the utility of CES-D as screener of depression (employing a cut-off score of 16) was studied, the incomplete versions gave somewhat less favourable results. As the rate of missing values increased, the diagnostic accuracy decreased.

At this point it may be informative to remind the reader that uncertainty must be distinguished from bias. Statistical uncertainty relates to the uncertainty with which population parameters are estimated from a sample. For example, parameter estimates of smaller samples are inclined to deviate more from the true values than of larger samples. By comparison, if such deviances result in systematical underestimation or overestimation of the outcomes then the estimates are said to be biased. Although the observed data were informative of the unobserved data, some imputation error was added to the CES-D total scores. However, these total scores did not deviate systematically from the complete data total scores. The extra uncertainty in total scores presents no problem when used as a measure of degree of mental health. As demonstrated by Graham et al. (1996) and Raghunathan and Grizzle (1995), the estimates of correlations between incomplete 'measure of degree' variables and other variables will not be biased; only the estimated standard errors will be larger. In contrast, the present study showed that the extra uncertainty in the total scores became bias when total scores were dichotomized on the basis of a cut-off score. Specifically, the imputation error clearly resulted in more misclassifications, decreasing the criterion validity. Moreover, diagnostic misclassifications can obscure the relationship of a diagnostic variable with other variables (see, e.g. Höfler, 2005). If the main aim of a research project is to give unbiased estimates of the relationship between a diagnosis and other variables, structurally incomplete designs may not be an optimal option. Then, all the effort should be put in the construction of an excellent, completely observed, screening instrument.

In contrast, if the estimation of such associations is not the main goal of a study, then incomplete designs may still be very useful. Given the outcomes of this study, the inaccuracy resulting from the incomplete design applied to the CES-D may be predicted from the rate of unobserved values of the design. On the basis of the expected accuracy, an appropriate design may be chosen. The acceptability of this inaccuracy depends on the research goal. For example, if a researcher wishes to detect depressed subjects, the sensitivity of the test should not be too low (e.g. minimally 60%). Therefore,

DOI: 10.1002/mpr

a design with a third missing (M1/3) may then be used at most. In contrast, if a researcher wishes to select non-depressed patients from the sample for further research, then the M3/4 design may be used. In the M3/4 design, in which instead of 20 only five items are administered per respondent, the negative predictive value merely drops 1 percentage point.

In the present study, the effect of incomplete designs was studied in a data file with a large number of observations. Of course, in many research situations much smaller numbers of observations are obtained. In smaller samples, the estimation of the relationships among the variables of incomplete data is more susceptible to sampling error, and therefore the estimation of the missing values may be less precise. However, this relationship between number of observations and precision is not limited to the analysis of missing data; it is true for statistical inference in general. Consequently, there is no reason to restrict the application of structurally incomplete designs to large samples. In addition to the number of observations, the statistical uncertainty associated with missing data is influenced by two things (e.g. Schafer, 1997, p. 61). First, the rate of missing values. In this study it was evident that the uncertainty increased as the rate of missing values of the design increased. Second, the strength of the relationships among the items. If the items of a test share more information, then the missing item scores are more easily predicted from observed item scores, which results in a lower statistical uncertainty of these estimates. Clearly, the statistical uncertainty is higher for test data with fewer observations, with an incomplete design that records fewer items per observation, and with items that have lower inter-item correlations.

It has been shown that in order to obtain a proper MI estimate of the relationship between variables suffering from missing values and a criterion variable, the criterion should be entered in the prediction model. If such a variable is left out, the criterion validity will be generally underestimated (see, e.g. Little, 1992; Schafer and Olsen, 1998). Therefore, for an unbiased estimate of the relationship between the incomplete data CES-D classification and DIS diagnosis, the DIS should have been included in the imputation model. The variable was not entered in the model because it was not our aim to validate the CES-D. Moreover, we mimicked common research settings in which diagnostic interviews are not administered (and therefore unavailable) because they are too costly, but instead a screener such as the CES-D is used. As a result of this exclusion of the DIS, the estimates of the diagnostic accuracy of the incomplete CES-D may have been too low.

In the simulation, the item sets were compiled ignoring the content of the items. In some situations however, it may be useful to take into account the item content to construct subsets. For example, when mental health scales contain items associated with multiple symptomatological dimensions. When each of these dimensions is equally represented in all subsets, then it is ensured that the uncertainty due to the incomplete design is evenly divided over these dimensions.

It should be understood that when applying DA, missing values need to be *multiply* imputed. Imputing a single data file and analysing it as if no missing values occurred will in all likelihood lead to spurious results. First, the estimates of parameters will be not be as precise as when averaging over multiple data sets because imputation error is not cancelled out. Second, standard errors will be biased because parameter uncertainty will not be properly reflected.

We studied the effect of applying incomplete designs to the CES-D in its role as both a screener of depression and a measure of degree of depression. However, mental health researchers may be interested in the development of depression as well. In such longitudinal research, structurally incomplete designs can also be convenient for data collection. How incomplete designs can be optimally used in such research settings must be studied in future research.

As for educational and psychological testing, our study showed that applying structurally incomplete designs may be useful in mental health testing. The proposed approach is a good alternative to the procedure suggested by Burisch (1997) and others, which employs item selection to make shorter tests. The difference between the two procedures is best described when picturing their application on a potentially complete data file. The test shortening procedure results in completely unobserved scores on the items that are not selected, and completely observed scores on the other items. The structurally incomplete test design results in a data file in which for each item a part is observed and a part is missing. Both procedures may cause the same rate of missing values in the data file. However, the incomplete test design has at least two advantages over the item selection procedure. First, information on more variables can be gathered. Second, because in the incomplete design all item combinations are observed,

DOI: 10.1002/mpr

the missing item scores can be estimated from the data; the result is that data on all items for all subjects are available. Of course, these two strategies are not mutually exclusive; they may be combined to come to an even shorter administration of a test. Redundant items may be removed using item selection methods, and to the shorter version an incomplete design may be applied when the data are gathered. Obviously, when this concerns mental health screeners, cut-off scores need to be adjusted.

Clearly, when applying incomplete designs, the responder's burden decreases. However, this does not come without a cost as the researcher has to do more work. More booklets need to be created. In addition, the predictive distribution of the missing values must be studied, and the data have to be multiply imputed and analysed.

This paper showed that missing data no longer exclusively create a problem, but also may provide a solution in many mental health research settings. When missing values result from carefully constructed incomplete designs, which cause the data to be missing completely at random, then sound estimates of these unobserved scores can be obtained with advanced missing data methods. The incomplete designs allow for the administration of a smaller set of items per respondent. This decreases the respondent burden, which may produce a higher quality of the items that are administered, and may even result in a higher response rate. However, compared to complete data, the incomplete designs may also introduce more statistical uncertainty, and therefore more diagnostic inaccuracy. Researchers may take into account both the total number of items to be recorded, the anticipated sample size, and the desired diagnostic accuracy to come to a design with a suitable rate of missing values.

References

Allison PD. Missing Data. Thousand Oaks, CA: Sage, 2001.

Beekman ATF, Deeg DJH, Smit JH, van Tilburg W. Predicting the course of depression in the elderly: results from a community-based study in the Netherlands. J Affect Disorders 1995; 34: 41–49.

Beekman ATF, Deeg DJH, van Limbeek J, Wouters L, van Tilburg W. Screening for depression in the elderly in the community: using the Center for Epidemiologic Studies Depression scale (CES-D) in the Netherlands. Tijdschr Gerontol Geriatr 1994; 25: 95–103.

Beekman ATF, Deeg DJH, van Limbeek J, Braam AW, de Vries MZ, van Tilburg W. Criterion validity of the Center for Epidemiologic Studies Depression scale (CES-D): results from a community-based sample of older subjects in the Netherlands. Psychol Med 1997; 27: 231–235. DOI: 10.1017/S0033291796003510

Bernaards CA, Sijtsma K. Factor analysis of multidimensional polytomous item response data suffering from ignorable item nonresponse. Multivar Behav Res 1999; 34: 277–313. DOI: 10.1207/S15327906MBR3403_1

Burisch M. Test length and validity revisited. Eur J Personality 1997; 11: 303–315.

Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas 1960; 20: 37–46.

Deeg DJH, Knipscheer CPM, van Tilburg W. Autonomy and well-being in the aging population: Concepts and design of the Longitudinal Aging Study Amsterdam, NIGTrend Studies No. 7. Bunnik, NL: Netherlands Institute of Gerontology, 1993.

Dillman DA, Sinclair MD, Clark JR. Effects of questionnaire length, respondent-friendly design, and a difficult question on response rates for occupant-addressed census mail surveys. Public Opin Quart 1993; 57: 289–304.

Dingemans P, van Engeland H, van Dijkhuis JH, Bleeker J. The diagnostic interview schedule (DIS). Tijdschr Psychiat 1985; 27: 341–359.

Enders CK. A primer on maximum likelihood algorithms available for use with missing data. Struct Equ Modeling 2001; 8: 128–141. DOI: 10.1207/S15328007SEM0801_7

Graham JW, Hofer SM, MacKinnon DP. Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. Multivar Behav Res 1996; 31: 197–218.

Hertzog C, van Alstine C, Usala PD, Hultsch DF, Dixon R. Measurement properties of the center for epidemiological studies depression scale (CES-D) in older populations. Psychol Assessment 1990; 2: 64–72.

Himmelfarb S, Murrell SA. Reliability and validity of five mental health scales in older persons. J Gerontol 1983; 38: 333–339.

Höfler M. The effect of misclassification on the estimation of association: a review. Int J Method Psych 2005; 14: 92–101. DOI: 10.1002/mpr.20

Johnson EG. The design of the national assessment of educational progress. J Educ Meas 1992; 29: 95–110.

Jongenelis K, Pot AM, Eisses AMH, Gerritsen DL, Derksen M, Beekman ATF, et al. Diagnostic accuracy of the original 30-item and shortened versions of the Geriatric Depression Scale in nursing home patients. Int J Geriatr Psych 2005; 20: 1067–1074. DOI: 10.1002/gps.1398

Kaplan D. The impact of BIB spiraling-induced missing data patterns on goodness-of-fit tests in factor analysis. J Educ Behav Stat 1995; 20: 69–82.

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159–174.

Little RJA. Regression with missing X's: a review. J Am Stat Assoc 1992; 87: 1227–1237.

Little RJA, Schenker N. Missing data. In Arminger G, Clogg CC, Sobel ME (eds) Handbook of Statistical Modeling for the Social and Behavioral Sciences, pp. 39–75. New York: Plenum Press, 1995.

May WL, Johnson WD. The validity and power of tests for equality of two correlated proportions. Stat Med 1997; 16: 1081–1096.

Mislevy RJ, Wu P-K. Missing Responses and IRT Ability Estimation: Omits, Choice, Time Limits, and Adaptive Testing, Technical Report No. RR-96-30-ONR. Princeton, NJ: Educational Testing Service, 1996.

Radloff LS. The CES-D scale: a self-report depression scale for research in the general population. Appl Psych Meas 1977; 1: 385–401.

Raghunathan TE, Grizzle JE. A split questionnaire survey design. J Am Stat Assoc 1995; 90: 54–63.

Robins L, Helzer JE, Croughan J, Radcliff KS. National Institute of Mental Health Diagnostic Interview Schedule: its history, characteristics and validity. Arch Gen Psychiatry 1991; 38: 381–389.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: Wiley, 1987.

Schafer JL. Analysis of Incomplete Multivariate Data. London: Chapman and Hall, 1997.

Schafer JL. NORM: Multiple Imputation of Incomplete Multivariate Data under a Normal Model, version 2, 1999. Available: http://www.stat.psu.edu/~jls/misoftwa.html

Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002; 7: 147–177. DOI: 10.1037//1082-989X.7.2.147

Schafer JL, Olsen MK. Multiple imputation for multivariate missing data problems: a data analyst's perspective. Multivar Behav Res 1998; 33: 545–571. DOI: 10.1207/s15327906mbr3304_5

Sijtsma K, van der Ark LA. Investigation and treatment of missing item scores in test and questionnaire data. Multivar Behav Res 2003; 51: 505–528. DOI: 10.1207/s15327906mbr3804_4

Smits N, Mellenbergh GJ, Vorst HCM. Alternative missing data techniques to grade point average: imputing unavailable grades. J Educ Meas 2002; 39: 187–206.

Tang WK, Wong E, Chiu HFK, Lum CM, Ungvari GS. The Geriatric Depression Scale should be shortened: results of Rasch analysis. Int J Geriatr Psych 2005; 20: 783–789. DOI: 10.1002/gps.1360

Wacholder S, Carroll RJ, Pee D, Gail MH. The partial questionnaire design for casecontrol studies. Stat Med 1994; 13: 623–634.

Correspondence: Niels Smits, Department of Clinical Psychology, Faculty of Psychology and Education, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands.
Email: n.smits@psy.vu.nl